

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-053395

(43)Date of publication of application : 26.02.1999

(51)Int.Cl.

G06F 17/30

(21)Application number : 09-219300

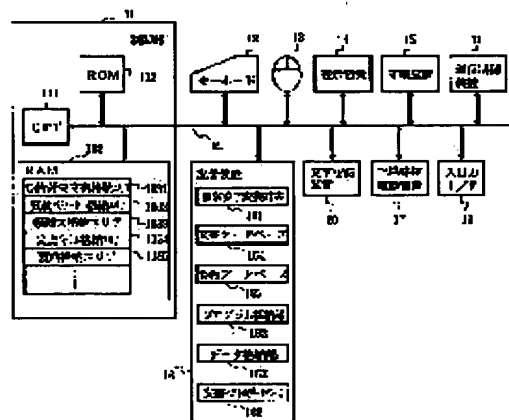
(71)Applicant : JUST SYST CORP

(22)Date of filing : 29.07.1997

(72)Inventor : NOMURA NAOYUKI  
FUJISAWA SHINJI**(54) DEVICE AND METHOD FOR DOCUMENT PROCESSING AND STORAGE MEDIUM  
STORING DOCUMENT PROCESSING PROGRAM****(57)Abstract:**

**PROBLEM TO BE SOLVED:** To extract the citation relation among plural documents and to produce the summaries including the proper conjunctions and connective sentences by adding the connective words among the summaries produced by a summary production means for the production of summaries of documents.

**SOLUTION:** A CPU 111 decides the importance of every candidate word (phrase) based on the summary parameters stored in a RAM 116 and the frequency of appearance, evaluation functions, etc., of extracted candidate words (phrases) set in every document group. Then the CPU 111 decides the importance of every sentence included in every document group based on the importance of every candidate word (phrase), the summary parameters, etc., and lists up the sentences within a summary ratio in the order of higher importance decided for the sentences. Furthermore, the CPU 111 arranges the listed-up sentences in the order of appearance set in the document groups to acquire a summary of the relevant document group. When the production of a summary is over to every document, the CPU 111 puts the conjunctions and connective sentences among the A produced summaries to produce a summary sentence.

**LEGAL STATUS**

[Date of request for examination]

28.07.2004

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

THIS PAGE BLANK (USPTO)

[Number of appeal against examiner's decision  
of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

**THIS PAGE BLANK (USPTO)**

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平11-53395

(43)公開日 平成11年(1999) 2月26日

(51)Int.Cl.<sup>6</sup>

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/401

3 2 0 A

15/40

3 7 0 A

審査請求 未請求 請求項の数10 F D (全 9 頁)

(21)出願番号 特願平9-219300

(22)出願日 平成9年(1997) 7月29日

(71)出願人 390024350

株式会社ジャストシステム

徳島県徳島市沖浜東3-46

(72)発明者 野村 直之

徳島県徳島市沖浜東3丁目46番地 株式会  
社ジャストシステム内

(72)発明者 藤澤 信二

徳島県徳島市沖浜東3丁目46番地 株式会  
社ジャストシステム内

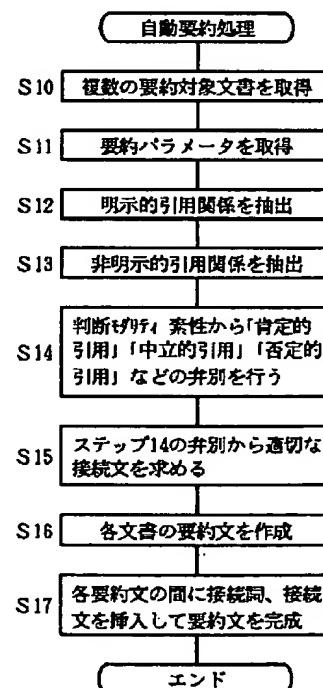
(74)代理人 弁理士 川井 隆 (外1名)

(54)【発明の名称】 文書処理装置、文書処理プログラムが記憶された記憶媒体および文書処理方法

(57)【要約】

【課題】 所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成する文書処理装置を提供すること。

【解決手段】 複数の文書間の関連性を判定するために、明示的引用関係と非明示的引用関係記述を抽出する。非明示的引用関係記述を抽出するためには、各毎の文書ベクトルを求め、各文書間で文書ベクトルの差をとる。これらの連続する2つの文書間のコサインバリュー (cosine value) が高いか低いかで非明示的引用関係があるか否かを判断する。そして、引用関係記述を肯定的引用、否定的引用、中立的引用に分け、接続文を生成する。そして、各文書を要約し、その文書間に接続文を挿入して、読み易い要約文を生成する。



**【特許請求の範囲】**

【請求項 1】 所定形式の文書を複数個取得する文書取得手段と、

前記文書取得手段により取得された複数の各文書間の関係を認識する関係認識手段と、

前記関係認識手段で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成手段と、

各文書の要約を自動的に作成する要約作成手段とを備え、

前記要約作成手段により作成された各要約間に前記接続語生成手段により生成された接続語を配置して複数文書の要約を作成することを特徴とする文書処理装置。

【請求項 2】 前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、

この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定手段をさらに備えたことを特徴とする請求項 1 記載の文書処理装置。

【請求項 3】 前記文書取得手段で取得された複数の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定手段を備え、

前記類似度判定手段は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定することを特徴とする請求項 2 記載の文書処理装置。

【請求項 4】 前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することを特徴とする請求項 2 または請求項 3 記載の文書処理装置。

【請求項 5】 所定形式の文書を複数個取得する文書取得機能と、

前記文書取得機能により取得された複数の各文書間の関係を認識する関係認識機能と、

前記関係認識機能で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成機能と、

各文書の要約を自動的に作成する要約作成機能とを備え、

前記要約作成機能により作成された各要約間に前記接続語生成機能により生成された接続語を配置して複数文書の要約を作成することをコンピュータに実現させるためのコンピュータ読取り可能な文書処理プログラムが記憶された記憶媒体。

【請求項 6】 前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、

この非明示的引用関係を検知するために各文書の類似性

を判断する類似度判定機能をさらに備えたことを特徴とする請求項 5 記載の文書処理プログラムが記憶された記憶媒体。

【請求項 7】 前記文書取得手段で取得された複数の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定機能を備え、

前記類似度判定機能は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定することを特徴とする請求項 6 記載の文書処理プログラムが記憶された記憶媒体。

【請求項 8】 前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することを特徴とする請求項 6 または請求項 7 記載の文書処理プログラムが記憶された記憶媒体。

【請求項 9】 所定形式の文書を複数個取得し、

取得された複数の各文書間の関係を認識し、

認識された関係に対応して各文書間に挿入する接続語を生成し、

各文書の要約を自動的に作成し、

作成した各要約間に生成した接続語を配置して複数文書の要約を作成することを特徴とする文書処理方法。

【請求項 10】 取得された複数の各文書間の関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、

この非明示的引用関係を検知するために各文書の類似性を判断することを特徴とする請求項 9 記載の文書処理方法。

**【発明の詳細な説明】****【0001】**

【発明の属する技術分野】この発明は、文書処理装置、文書処理方法および文書処理プログラムを記憶した記憶媒体に係り、詳細には、複数の文書から文書間の関係に言及した要約を作成する技術に関する。

**【0002】**

【従来の技術】従来、書籍、論文、報告書等の各種の文書に対し、要約（抄録を含む）の自動作成処理や、他文書等との関連づけ処理等の各種処理をコンピュータを用いて行うことが行われている。文書の自動要約については、例えば、「全文情報からの意味的情報の抽出と加工」（情報処理学会第 3 8 回全国大会予稿集、第 2 2 2 頁；1989 年）で提案されている。この方法では、まず文書中の重要語を字種や動詞等の情報から抽出し、さらに重要語の出現頻度から最重要語を決定する。次に、重要語と最重要語が出現するか否かから重要文を決定することで、自動的に要約を作成することが可能になる。また、文章の段落の性質を反映させることで、より正確

に要約を作成する特開平 3-191475 号公報に記載された方法等も提案されている。一方、他のデータとの関連づけとしては、インターネットにおけるハイパーリンクや、フレームシステム等による知識処理（エキスパートシステム等）における関連づけ等が行われている。

【0003】

【発明が解決しようとする課題】ところで、従来の文書処理装置では、単数の文書を対象として、要約するものであった。そこで、複数の文書について要約を作成する場合、個々の文書を要約してこれを繋ぎ合わせる必要であった。しかし、この方法によると、複数の各文書が同一のトピックのみで構成されている場合は、比較的適切な要約を作成することが可能であるが、各文書が異なる複数のトピックを含むときは、必ずしも適切な要約を作成することができなかった。すなわち、各文書の内容の類否を考慮せず、互いに異なる主張や事実の記載をもつ複数文書の要約を互いにつなぎ合わせることで要約を作成していたため、可読性の低い要約を生成していた。

【0004】そこで、本発明は、このような従来の課題を解決するためになされたもので、所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成する文書処理装置および文書作成方法を提供することを第 1 の目的とする。また、本発明は、所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成することができるコンピュータ読取り可能な文書処理プログラムを記憶した記憶媒体を提供することを第 2 の目的とする。

【0005】

【課題を解決するための手段】請求項 1 記載の発明では、文書処理装置が、所定形式の文書を複数個取得する文書取得手段と、前記文書取得手段により取得された複数の各文書間の関係を認識する関係認識手段と、前記関係認識手段で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成手段と、各文書の要約を自動的に作成する要約作成手段とを備え、前記要約作成手段により作成された各要約間に前記接続語生成手段により生成された接続語を配置して複数文書の要約を作成することにより前記第 1 の目的を達成する。

【0006】請求項 2 に記載した発明では、請求項 1 に記載した文書処理装置において、前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定手段をさらに備えたことにより前記第 1 の目的を達成する。

【0007】請求項 3 に記載した発明では、請求項 1 または請求項 2 に記載した文書処理装置において、前記文

書取得手段で取得された複数個の各文書の特徴づける文書ベクトルを決定する文書ベクトル決定手段を備え、前記類似度判定手段は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定する。

【0008】請求項 4 に記載した発明では、請求項 2 または請求項 3 記載の文書処理装置において、前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識する。

【0009】請求項 5 に記載した発明では、文書処理プログラムが記憶された記憶媒体が、所定形式の文書を複数個取得する文書取得機能と、前記文書取得機能により取得された複数の各文書間の関係を認識する関係認識機能と、前記関係認識機能で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成機能と、各文書の要約を自動的に作成する要約作成機能とを備え、前記要約作成機能により作成された各要約間に前記接続語生成機能により生成された接続語を配置して複数文書の要約を作成することを実現させることにより前記第 2 の目的を達成する。

【0010】請求項 6 に記載した発明では、請求項 5 記載の文書処理プログラムが記憶された記憶媒体に、前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定機能を実現させて前記第 2 の目的を達成する。

【0011】請求項 7 に記載した発明では、請求項 6 記載の文書処理プログラムが記憶された記憶媒体において、前記文書取得手段で取得された複数個の各文書の特徴づける文書ベクトルを決定する文書ベクトル決定機能を備え、前記類似度判定機能は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定させて前記第 2 の目的を達成する。

【0012】請求項 8 に記載した発明では、請求項 6 または請求項 7 記載の文書処理プログラムが記憶された記憶媒体において、前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することにより前記第 2 の目的を達成する。

【0013】請求項 9 に記載した発明では、所定形式の文書を複数個取得し、取得された複数の各文書間の関係を認識し、認識された関係に対応して各文書間に挿入する接続語を生成し、各文書の要約を自動的に作成し、作

成した各要約間に生成した接続語を配置して複数文書の要約を作成することにより前記第1の目的を達成する。

【0014】請求項10に記載した発明では、請求項9に記載した発明において、取得された複数の各文書間の関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断することにより前記第1の目的を達成する。

【0015】

【発明の実施の形態】以下、本発明の文書処理装置、文書処理方法および文書処理プログラムを記憶した記憶媒体の好適な実施の形態を、図1ないし図8を参照して詳細に説明する。

(1) 実施の形態の概要

本実施の形態では、複数の文書間の関連性を判定するために、明示的引用関係と非明示的引用関係記述を抽出する。非明示的引用関係記述を抽出するためには、各毎の文書ベクトルを求め、各文書間で文書ベクトルの差をとる。これらの連続する2つの文書間のコサインバリュウ(cosine value)が高いか低いかで非明示的引用関係があるか否かを判断する。そして、引用関係記述を肯定的引用、否定的引用、中立的引用に分け、接続文を生成する。そして、各文書を要約し、その文書間に接続文を挿入して、読み易い要約文を生成する。

【0016】(2) 実施の形態の詳細

図1は、文書処理装置の構成を表したブロック図である。本実施の形態の文書処理装置は、パーソナルコンピュータやワードプロセッサ等を含むコンピュータシステムとして構成し、また、LAN(ローカル・エリア・ネットワーク)のサーバーやインターネットを含むコンピュータ(パソコン)通信のホストとして構成することが可能である。文書処理装置は、図1に示すように装置全体を制御するための制御部11を備えている。この制御部11には、データバス等のバスライン21を介して、入力装置としてのキーボード12やマウス13、表示装置14、印刷装置15、記憶装置16、記憶媒体駆動装置17、通信制御装置18、および、入出力I/F19、および、文字認識装置20が接続されている。制御部11は、CPU111、ROM112、RAM113を備えている。ROM112は、CPU111が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。

【0017】RAM113は、CPU111にワーキングメモリとして使用されるランダム・アクセス・メモリである。このRAM113には、本実施の形態による要約処理を行うためのエリアとして、要約対象文書格納エリア1131、要約パラメータ格納エリア1132、接続文格納エリア1133、文書ベクトル格納エリア1134、要約格納エリア1135、その他の各種エリアが

確保されるようになっている。要約パラメータ格納エリア1132には、操作者からの入力等により取得された要約パラメータの値または後述のデータ格納部の163から読み込んだ要約パラメータのデフォルト値が格納される。操作者が入力する要約パラメータとしては、例えば、全文書に対する要約の比率(1%~99%)、数量優先のある/なし、長単文のある/なし、です/であるの選択をする/しない、等の値が格納される。接続文格納エリア1133には、各要約文を接続するために生成された接続文が格納される。文書ベクトル格納エリア1134には、要約対象文書に対する文書ベクトルと、後述する各サブ文書に対する文書ベクトルとが格納される。要約格納エリア1135には、各文書の要約文が格納される。

【0018】キーボード12は、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。マウス13は、ポインティングデバイスであり、表示装置14に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。表示装置14は、例えばCRTや液晶ディスプレイ等が使用される。この表示装置には、要約対象文書の内容や、本実施の形態により自動生成された要約の内容等が表示されるようになっている。印刷装置15は、表示装置14に表示された文章や、記憶装置16の文書格納部164に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットブリタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。

【0019】記憶装置16は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置16に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置17で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。記憶装置16は、仮名漢字変換辞書161、プログラム格納部162、データ格納部163、文書データベース164、要約データベース165、文書ベクトルデータベース166、図示しないその他の格納部(例えば、この記憶装置16内に格納されているプログラムやデータ等をバックアップするための格納部)等を有している。プログラム格納部162には、本実施の形態における自動要約処理プログラム、文書ベクトル作成処理プログラム、要約作成処理プログラム等の各種プログラムの他、仮名漢字変換辞書161を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。データ格納部163には、要約パラメータのデフォルト値等の各種データが格納され



ている。要約パラメータのデフォルト値としては、例えば、全文書に対する要約の比率＝「25%」や、日付時刻、価格情報、物理量（サイズ、重量、温度等）等の数量重視＝「しない」や、URL（Uniform Resource Locator）重視＝「しない」や、です／ます／であるの選択＝「しない」、等の値が格納されている。

【0020】文書データベース164には、仮名漢字変換プログラムにより作成された文書や、他の装置で作成されて記憶媒体駆動装置17や通信制御装置18から読み込まれた文書が格納される。この文書データベース164に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、HTML（Hyper Text Markup Language）形式の文書、JIS形式の文書等の各種形式の文書の格納が可能である。文書データベース164には、これらの形式の文書データの格納される。要約データベース165、及び文書ベクトルデータベース166には、文書データベース164に格納されている各文書に対応する要約や文書ベクトルが格納されるようになっている。

【0021】図2は、文書ベクトルデータベース166の内容を概念的に表したものである。この図2に示されるように、文書中から自動抽出されたキーワード $x$ に対して求められた要素値 $f(x)$ が文書ベクトルの要素として格納されている。この文書ベクトルは各文書（A、B、C…）毎に格納され、文書データベース164に格納されている各文書と対応づけられている。各文書ベクトルの次元は採用するキーワード $x$ （重要語句）の数であるが、2文書間の類似度を両文書ベクトルから求める場合には、両文書のキーワードの和集合の数が両文書ベクトルの次元となる。この場合、一方の文書ベクトルにのみ含まれるキーワードに対する他方の文書ベクトルの要素値は、“0”に定義される。

【0022】例えば、図2において、文書Bのキーワードは「重要、重要語、重要度、…」、文書Cのキーワードは「重要、…、政治、…」であり、両文書の文書ベクトルは次の通りである。

文書Bの文書ベクトル＝（ 1, 18, 19, …）

文書Cの文書ベクトル＝（18, …, 21, …）

これに対して文書Bと文書Cとの類似度を算出する場合には、両文書のキーワードを「重要、重要語、重要度、…、政治、…」とし、両文書の文書ベクトルはつぎの通り定義される。

文書Aの文書ベクトル＝（ 1, 18, 19, …, 0, …）

文書Cの文書ベクトル＝（18, 0, 0, …, 21, …）

【0023】記憶媒体駆動装置17は、CPU111が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラムには、本実施

の形態の文書処理装置により実行される各種処理のためのプログラム、および、そこで使用される辞書、データ等も含まれる。ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやICカード等の半導体記憶媒体、CD-ROMやMO、PD（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。本実施の形態の文書処理装置において使用される記憶媒体としては、主として、CD-ROMやフロッピーディスクが使用される。記憶媒体駆動装置17は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み可能な記憶媒体に対してRAM113や記憶装置16に格納されているデータ等を書き込むことが可能である。

【0024】本実施の形態の文書処理装置では、制御部11のCPU111が、記憶媒体駆動装置17にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置16の各部に格納する。そして、本実施の形態による自動要約処理等の各種処理を実行する場合、記憶装置16から該当プログラムをRAM113に読み込み、実行するようになっている。但し、記憶装置16からではなく、記憶媒体駆動装置17により外部の記憶媒体から直接RAM113に読み込んで実行することも可能である。また、文書処理装置によっては、本実施の形態の自動要約処理プログラム等を予めROM112に記憶しておき、これをCPU111が実行するようにしてもよい。

【0025】通信制御装置18は、他のパーソナルコンピュータやワードプロセッサ等との間でテキスト形式やHTML形式等の各種形式の文書やビットマップデータ等の各種データの送受信を行うことができるようになっている。入出力I/F19は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。文字認識装置20は、用紙等に記載された文字をテキスト形式やHTML等の各種形式で認識する装置であり、イメージスキャナや文字認識プログラム等で構成されている。

【0026】本実施の形態では、キーボード12の入力操作により作成した文書（RAM113の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置17から読み込んだ文書、予め文書データベースに格納されている文書、通信制御装置18からダウンロードした文書、及び文字認識装置20で文字認識した文書、等の各種文書を対象文書として取得する（文字取得手段）ことが可能である。

【0027】以上のように構成された本実施の形態の文書処理装置による、複数文書から要約を作成する自動要約処理の動作について図3から図8を用いて説明する。図3は自動要約処理のメイン動作を表したものである。図4中に示した文書ベクトルは、概念的に理解しやすくするために2次元で表示したものであるが、実際にはN次元ベクトルである。CPU111は、要約を作成する対象となっている複数の文書の1つである要約対象文書A(図4(A))を取得し、RAM113の要約対象文書格納エリア1131に格納する(ステップ10)。要約対象文書は、ユーザの指示に従ってRAM113(自装置内で作成された文書である場合)、記憶装置16の文書データベース164(要約が未だ作成されていない文書である場合)、記憶媒体駆動装置17(自装置または他装置で作成済みの文書の場合)、通信制御装置18(パソコン通信、インターネット等の通信による場合)から取得する。

【0028】次に、CPU111は、ユーザによってキーボード12等から要約パラメータが入力された場合には入力値を取得し、ユーザによる入力がない場合にはデータ格納部163に格納された要約パラメータのデフォルト値を取得し、要約パラメータ格納エリア1132に格納する(ステップ11)。

【0029】次に、取得した各文書間における明示的引用関係を抽出する(ステップ12)。ここで、明示的引用関係とは、例えば「前記…」、「…という見解がある」、「…著」など文書の外見から明白に他の文書を引用していることが明らかな引用をいう。この抽出は、特約語、特約表現知識ベースを用いて形態素解析にマッチングをかけて行う。続いて、取得した各文書間における非明示的引用関係を抽出する(ステップ13)。非明示的引用関係とは、テキスト内容で同一のイベントに言及しつつ「…という見解もあるようだが、」のように「…」で纏められたテキストの中から、より古い記事側との類似性を判定して得られる。この文書間の類似性を

類似度  $s = \cos(\alpha)$

$$= (b_n \cdot b_{n+1}) / (|b_n| \times |b_{n+1}|)$$

【0033】この類似度  $s$  の値は  $-1 \leq s \leq 1$  までの値をとり、1に近いほど2つの文書ベクトルが互いに平行に近く、2つの文書同士は似ていると考えることができる。図5に示すように、AとBとのようにベクトルの向きが近似しているものは、内容が近似しており、A、BとCとのように、ベクトルの向きが異なるものは、内容が相違していることとなる。その後、ステップ12、ステップ13で抽出した引用関係の記述の近傍にある筆者の判断のモダリティ素性から「肯定的引用」「中立的引用」「否定的引用」などの弁別を行う(ステップ14)。筆者の判断のモダリティ素性とは、例えば、「遺憾ながら…間違っている」「賛成できない」「反対せざるを得ない」などの記述を基に決定する。ステップ14

判定するために、CPU111は、要約対象文書格納エリア1131に格納した要約対象文書の各文章に対する文書ベクトルV(図4)を求める。

【0030】図7は、文書ベクトル作成処理の動作を表したフローチャートである。CPU111は、形態素解析を行うことで要約対象文書の文章から自立語を抽出する(ステップ131)と共に、名詞句、複合名詞句等を含めた候補語(句)を要約対象文書Aから抽出しRAM113の所定作業領域に格納する(ステップ132)。そして抽出した候補語(句)の要約対象文書での出現頻度、評価関数から、各候補語(句)重要度  $f(x)$  を決定する(ステップ133)。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語(句)の種類による重み付け等が使用される。さらにCPU111は、決定した重要度  $f(x)$  の値から要約対象文書Aのキーワード  $a, b, \dots$  を決定する(ステップ134)。そして、各キーワードの重要度  $f(x)$  を要素として、文書ベクトル  $V = (f(a), f(b), \dots)$  をRAM113の文書ベクトル格納エリア1134に格納する(ステップ135)。この文書ベクトルVを求める処理を複数の各文書B、C、D……と全ての要約対象文書について行う。

【0031】要約対象の全ての文書に対して文書ベクトルVが求まるとCPU111は、各文書間の類似度を求める(ステップ13)。各文書間の類似度  $s$  を、両者の文書ベクトル  $b_n$  と文書ベクトル  $b_{n+1}$  間の角度に依存するコサインにより求める。すなわち、両文書ベクトル  $b_n$  と  $b_{n+1}$  間の角度を  $\alpha$  とし、両文書ベクトルの内積を  $b_n \cdot b_{n+1}$  とし、両文書ベクトルの大きさをそれぞれ  $|b_n|$ 、 $|b_{n+1}|$  とした場合、両文書ベクトルの類似度  $s$  は次の数式1により求まる。

【0032】

【数1】

の弁別から各々接続詞と、トピック文、文末表現からなる「接続文」を生成する(ステップ15)。例えば、「同様に、鈴木氏が文献Aでも『〇〇〇』と記述している。」「しかし、一方では、田中氏による『…文献A』という意見もある。」「また、…(木村三郎氏：文献B)との報告がある。」「の各文の「同様に」「しかし」「また」といった接続詞、「と記述している。」「という意見もある。」「との報告がある。」「との「接続文」を用意する。

【0034】そして、各文書の要約文を作成する(ステップ16)。図8は、要約作成処理の動作を表したフローチャートである。CPU111は、まず形態素解析を行うことで各文書群に含まれる自立語を抽出する(ステ

ップ221)と共に、名詞句、複合名詞句等を含めた候補語(句)を要約対象文書Aから抽出しRAM113の所定作業領域に格納する(ステップ222)。そして、RAM16の要約パラメータ格納エリア1132に格納した要約パラメータや、抽出した候補語(句)の各文書群中での出現頻度、評価関数等から、各候補語(句)の重要度 $f(y)$ を決定する(ステップ223)。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語(句)の種類による重み付け等が使用される。

【0035】さらにCPU111は、決定した重要度 $f(y)$ や要約パラメータ格納エリア1132に格納された要約パラメータ等から、各文書群に含まれる各センテンスに対する重要度 $F(z)$ を決定する(ステップ224)。そして、決定したセンテンスの重要度 $F(z)$ の重要度が高いセンテンスの上位から要約パラメータの要約比率(例えば、文書群の全センテンス数の内の上位25%)以内に入るセンテンスをリストアップする(ステップ225)。そしてCPU111は、リストアップしたセンテンスを文書群の中での出現順に並べることで当該文書群についての要約とし、これをRAM113の要約格納エリアに格納して(ステップ226)、図3の自動要約処理ルーチンにリターンする。

【0036】各文書に対する要約の作成が終了するとCPU111は、図6に示すように、作成された要約文の間にステップ15で生成した接続詞、接続文を挿入して、要約文を完成する(ステップ16)。これを要約格納エリア1135の所定エリアに格納して、本実施の形態による自動要約処理を終了する。以上説明したように、本実施の形態による自動要約処理によれば、複数の文書にまたがり、且つそれらの間の関係についての短い説明文を含む、可読性の高い要約を自動生成することができる。

【0037】以上の自動要約処理が終了すると、CPU111はユーザの指示によりRAM113に格納した各データの保存処理を行う。すなわち、要約対象文書格納エリア1131から要約対象文書を読み出して、記憶装置16の文書データベース164に格納する。また作成した要約を要約格納エリア1135から読み出し、文書データベース164に格納した要約対象文書との関連性を付けて記憶装置16の要約データベース165に格納する。さらに、文書ベクトル作成処理で求めた文書ベクトルVを文書ベクトル格納エリア1135から読み出し、文書データベース164に格納した要約対象文書との関連性を付けて記憶装置16の文書ベクトルデータベース166に格納する。

【0038】以上、本実施の形態の構成および自動要約処理について説明したが、本発明では、これらの各形態に限定されるものではなく、請求項に記載された発明の

範囲内で種々の変形をすることが可能である。例えば実施の形態では、形態素解析及び候補語(句)の抽出について、文書ベクトル作成処理(図7のステップ131とステップ132)と、要約作成処理(図8のステップ221とステップ222)とにおいて独立して同様な処理を行うこととしたが、本発明では、文書ベクトル作成処理で抽出した候補語(句)をRAM16の所定エリアに格納しておき、要約作成処理で利用するようにしてもよい。

【0039】また説明した実施の形態では、自動要約処理が終了した後の保存処理において、要約対象文書、要約、文書ベクトルVのみを記憶装置16の各データベース164、165、166に格納し保存するようにしたが、本発明では更に、文書ベクトル作成処理(図7)のステップ132で要約対象文書から抽出し、RAM113の所定作業領域に格納した候補語(句)を要約対象文書Aと関連付けて、文書データベース164、又は専用の候補語(句)データベースに格納するようにしてもよい。また要約パラメータ格納エリア1132から要約パラメータを読み出して、当該要約に関連付けて、要約データベース166、または専用の要約パラメータデータベースに格納するようにしてもよい。

【0040】さらに、説明した実施の形態では、文書ベクトル作成処理(ステップ13、図7)及び要約作成処理(ステップ22、図8)の両処理において、形態素解析(ステップ131、221)と候補語(句)の抽出(ステップ132、222)を行った。しかし、同一センテンスに対する処理であるため、抽出した候補語(句)は同一である。そこで、本発明では、文書ベクトル作成処理で抽出した候補語(句)をRAM113の所定エリアに格納しておき、要約処理において格納した候補語(句)を使用することでステップ221とステップ222を省略するようにしてもよい。この候補語(句)についても、要約対象文書に対する候補語(句)として文書データベース164、又は専用の候補語(句)データベースに格納するようにしてもよい。

【0041】説明した実施の形態では文書ベクトルを作成する方法として図7のフローチャートに従った方法を1例にして説明したが、本発明でこの方法に限られるものではなく、要約対象文書中Aからキーワードを抽出する方法や、抽出キーワードに対する重要度(=文書ベクトルの要素値)の決定方法等については、公知の各種方法により置き換えることが可能である。また、各サブ文書群に対する要約の作成処理についても同様に図8のフローチャートに示した方法に限られるものではなく、公知の各種要約方法、抄録作成方法等を資料することが可能である。更に、2つの文書ベクトルの類似度の算出方法については、数式1により類似度を算出することとしたが、この数式に限定されるものではなく、ベクトル相互間の類似関係を表すことが可能であれば他の数式によ

り類似度を算出することも可能である。

【0042】説明した実施の形態では、日本語で作成された文書に限られるものでなく、あらゆる言語で作成された文書を対象とすることが可能である。その場合、対象となる文書が作成された言語用の形態素解析アルゴリズム等を使用するといった、本発明の構成には影響のない部分を変更するだけでよい。なお、以上の実施の形態において説明した、各装置、各部、各動作、各処理等に対しては、それらを含む上位概念としての各手段（～手段）により、実施の形態を構成することが可能である。例えば、「決定した重要度  $f(x)$  の値から要約対象文書 A のキーワード  $a, b, \dots$  を決定する（ステップ 134）」との記載に対して「キーワード決定手段」を構成し、「決定したセンテンスの重要度  $F(z)$  の重要度が高いセンテンスの上位から要約パラメータの要約比率（例えば、サブ文書群の全センテンス数の内の上位 25%）以内に入るセンテンスをリストアップする（ステップ 225）」との記載に対して「センテンスリストアップ手段」を構成するようにしてもよい。同様に、その他各種動作に対して「～（動作）手段」等の上位概念で実施の形態を構成するようにしてもよい。

#### 【0043】

【発明の効果】本発明によれば、複数の文書で構成された所定形式の文書を取得し、取得した文書との関係を引用関係から把握して、適切な接続文を生成し、これを要約した各要約文の間に挿入することで、前後の文脈が明確な分かり易い複数文書の要約を得ることが出来る。

#### 【図面の簡単な説明】

【図 1】本発明の 1 実施の形態における文書処理装置の構成を表したブロック図である。

【図 2】同上、実施の形態における文書ベクトルデータベースの内容を概念的に表した説明図である。

【図 3】同上、実施の形態における自動要約処理のメイン動作を表したフローチャートである。

【図 4】同上、実施の形態における、文書 A に対する文書ベクトルを求めたところ示す図である。

【図 5】同上、実施の形態における、各文書に対する文書ベクトルを求めたところ示す図である。

【図 6】同上、実施の形態における、要約文と接続文の接合を示す図である。

【図 7】同上、実施の形態における文書ベクトル作成処理の動作を表したフローチャートである。

【図 8】同上、実施の形態における要約作成処理の動作を表したフローチャートである。

#### 【符号の説明】

- 11 制御部
- 112 ROM
- 113 RAM
- 1131 要約対象文書格納エリア
- 1132 要約パラメータ格納エリア
- 1133 接続文格納エリア
- 1134 文書ベクトル格納エリア
- 1135 要約格納エリア
- 12 キーボード
- 13 マウス
- 14 表示装置
- 15 印刷装置
- 16 記憶装置
- 161 仮名漢字変換辞書
- 162 プログラム格納部
- 163 データ格納部
- 164 文書データベース
- 165 要約データベース
- 166 文書ベクトルデータベース
- 17 記憶媒体駆動装置
- 18 通信制御装置
- 19 入出力 I/F
- 20 文字認識装置

【図 2】

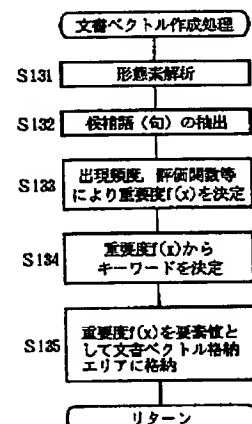
文書ベクトルデータベース

文 書	キーワードの要素値 $f(x)$					
	.....	重 要	重要語	重要度	.....	政 治
A	.....	2	20	21	.....	—
B	.....	1	18	19	.....	—
C	.....	18	—	—	.....	21
.....	.....	.....	.....	.....	.....	.....

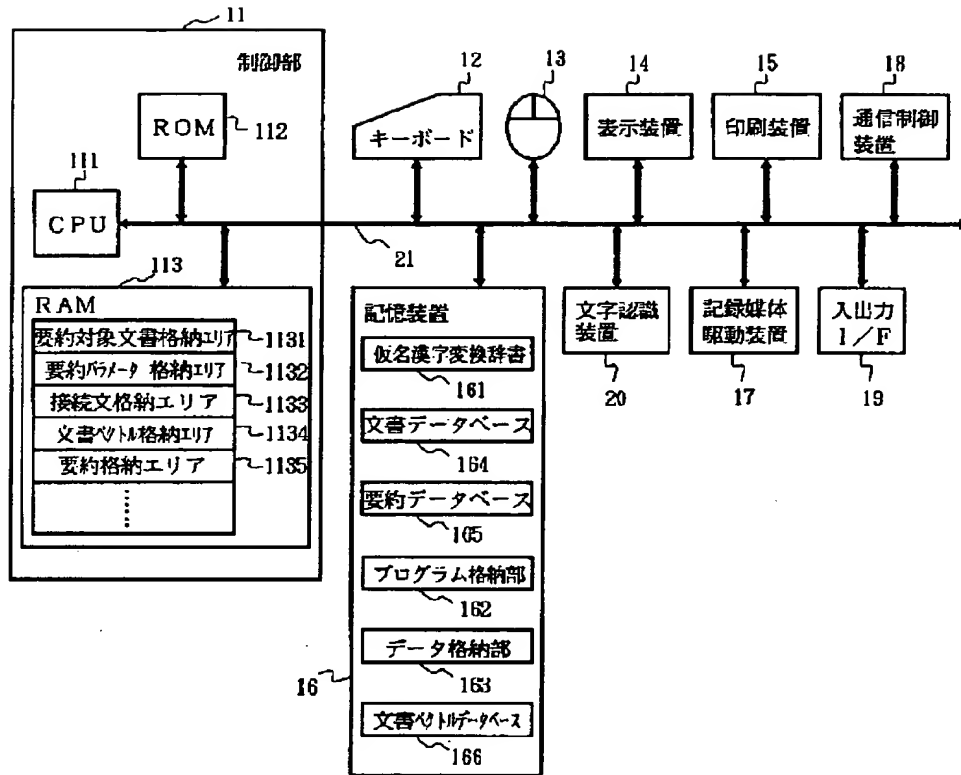
【図 5】

文書 A	文書ベクトル $b$ を求める	
	$b_1$	$b_2$
文書 B	$b_2$	$b_3$
文書 C	$b_3$	$b_4$
⋮	⋮	⋮

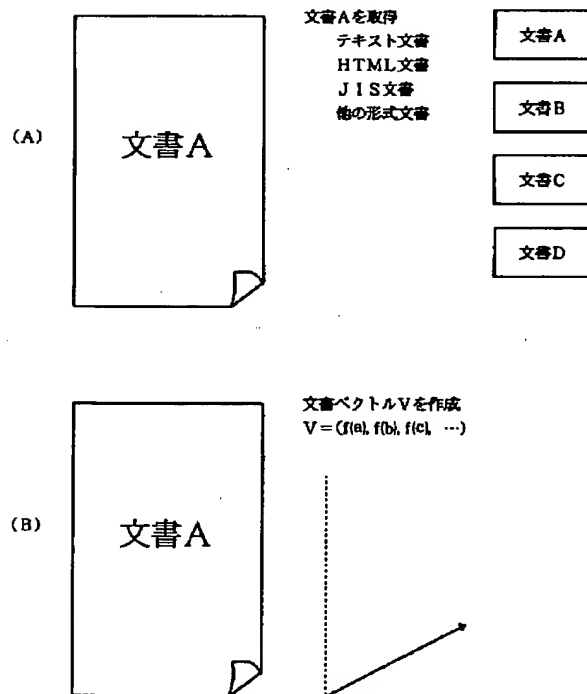
【図 7】



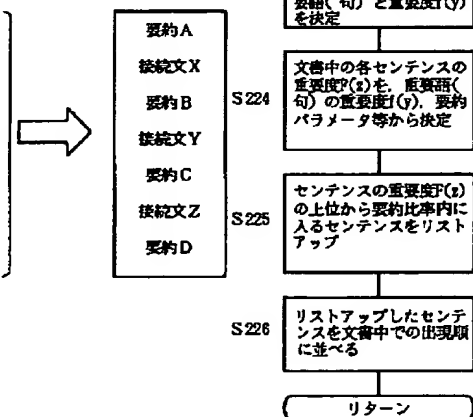
【図 1】



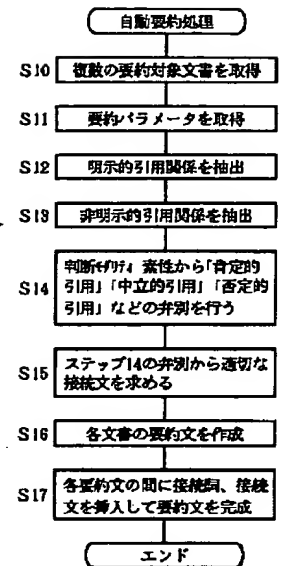
【図 4】



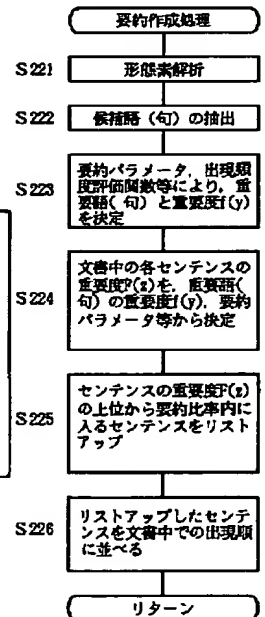
【図 6】



【図 3】



【図 8】



THIS PAGE BLANK (USPTO)